Illuminating the Scope and Impact of Open Source Software: A Framework for Measurement with an Application Based on the R Program and its Impacts

Carol A. Robbins
NSF, National Center for Science and Engineering Statistics
crobbins@nsf.gov

Gizem Korkmaz
Virginia Tech, Social and Decision Analytics Lab

Claire Kelling
Penn State University

Open source software is everywhere, both as specialized applications nurtured by devoted user communities, and as digital infrastructure underlying platforms used by millions daily. This type of software is developed, maintained, and extended both within the private sector and outside of it, through the contribution of people from universities, government research institutions, nonprofits, and individuals. Examples include Linux, Apache, Python, and R. Despite its ubiquity and extensive use, while GDP measures generally account for business sector software as an intangible asset or intellectual property product, reliable measures of the scope and impact of this type of software outside of the business sector are scarce. Open source software is a component of digital dark matter, products with the characteristics of public goods and private goods that serve as inputs to production as well as sources of non-pecuniary and effectively limitless benefits (Greenstein and Nagle, 2013). This digital dark matter includes activities currently missing or not well measured in existing innovation statistics, characterized as dark innovation. Martin (2015) describes this as including incremental innovation activities, innovations in non-manufactured products, innovations that take place without formal R&D, and innovations that are unpatented. The implementation of open source software spans all four of these activities.

While the extent and impact of open source software is currently unknown, recent estimates suggest that its magnitude is significant. Greenstein and Nagle (2013) estimate that the Apache server, developed initially at the National Center for Supercomputing Applications at the University of Illinois, is equivalent to between 1.3 and 8.7 percent of the stock of prepackaged software currently accounted for in US private fixed investment. In this paper, we present a framework for measurement of production and use of open source software based on a sectoral analysis for the US: universities, government institutions, nonprofit organizations, households, and businesses. Our goal is to estimate the contribution of universities and government institutions to investments in open source software. This information is increasingly important for understanding the full scope of research outputs, which are most frequently estimated using patents, publications, and citations. While current and comprehensive survey data do not exist for this the contributions of open source software, because open source software is disseminated online, a wealth of information is available to be scraped, both metadata and information embedded in repositories and in the code and headers of the software programs themselves. Our contribution is to show how these data can be used to develop measures of open source software that are complementary to GDP measures of private investment, shedding light on sources of innovation and productivity growth currently not well understood.

In addition to this framework for sectoral analysis, we present an analysis of factors that affect the impact of open source software, focusing on a case study of R packages. R is a programming language R was created in 2000 at the University of Auckland. It is one of the fastest growing programming languages, widely used for data mining, data analysis, visualization, and statistical modeling. Thus far, we have collected all the R packages hosted on the repository CRAN, and on depsy.org, which is a website that compiles R and Python packages. To date, information about 9,810 R packages have been scraped from Depsy.org (last update in Sept. 2015). The information for each package includes its contributors, number of commits, number of downloads, number of citations, and stars (identifying active development). To estimate the impact of R packages, we develop two Quasi-Poisson models with the number of downloads (Model I) and the number of citations (Model II) as the dependent variables, and use the network characteristics and the package attributes as independent variables.

We find that the network centrality of a package (the number of packages that depend on it, and how they are connected) as well as package attributes (e.g., the number of contributors, the number of commits) are important factors in showing the impact of open source software, measured by the number of downloads and the number of citations. In our preliminary work, we have focused on complementarities and generate the dependency network of the packages, where a directed edge i &#8594; j indicates that the package j requires i to be installed to function. We obtain a network with 7,389 nodes and 20,235 directed edges. The average degree (indegree and outdegree) of the dependency network is found as 2.74. Our goal is to estimate the value of R from two perspectives, the cost of creation and an approximation of market value. The cost method estimates the total effort to develop a software package using information embedded within the code about complexity. The approximation of market value accounts for substitutes include proprietary software such as SAS and Stata, and complementarities include both software that build on R, such as Tableaux, a proprietary program for data visualization.

References:

Greenstein and Nagle, 2013. Digital Dark Matter and the Economic Contribution of Apache, NBER Working Paper 19507

Martin, Ben, 2015. Twenty Challenges for Innovation Studies, Science and Public Policy, Volume 43, Issue 3, 1 June 2016, Pages 432–450, https://doi.org/10.1093/scipol/scv077

Ver Hoef, J. M. and Boveng, P. L., 2007. Quasi-Poisson vs Negative Binomial Regression: How Should We Model Overdispersed Count Data?. Ecology, 88: 2766–2772. doi:10.1890/07-0043.1